3D Convolutional Neural Networks for Event-Related Potential detection

H. Cecotti *Senior Member, IEEE* and G. Jha

Abstract— **Deep learning techniques have recently been successful in the classification of brain evoked responses for multiple applications, including brain-machine interface. Single-trial detection in the electroencephalogram (EEG) of brain evoked responses, like event-related potentials (ERPs), requires multiple processing stages, in the spatial and temporal domains, to extract high level features. Convolutional neural networks, as a type of deep learning method, have been used for EEG signal detection as the underlying structure of the EEG signal can be included in such system, facilitating the learning step. The EEG signal is typically decomposed into 2 main dimensions: space and time. However, the spatial dimension can be decomposed into 2 dimensions that better represent the relationships between the sensors that are involved in the classification. We propose to analyze the performance of 2D and 3D convolutional neural networks for the classification of ERPs with a dataset based on 64 EEG channels. We propose and compare 6 conv net architectures: 4 using 3D convolutions, that vary in relation to the number of layers and feature maps, and 2 using 2D convolutions. The results support the conclusion that 3D convolutions provide better performance than 2D convolutions for the binary classification of ERPs.**

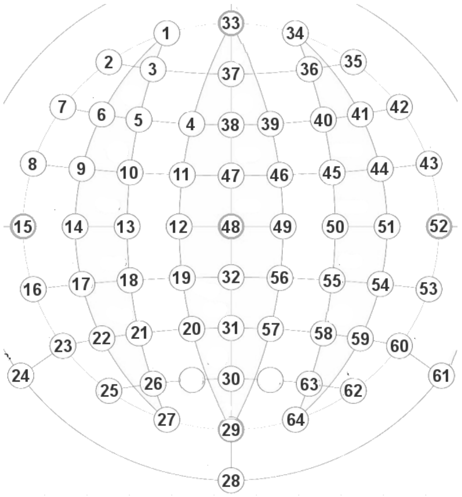
I. Introduction

The detection of brain responses using electroencephalogram (EEG) signals at the single-trial level has implications in brain-machine interface, as an increase of performance in the accuracy is synonym of an improvement of the information transfer rate between the brain and the machine. In addition, single-trial detection provides a tool for the analysis of the brain evoked responses dynamic over time. To achieve this task, machine learning and signal processing are required [1]. Classifiers based on deep learning approaches have gained a great success in biomedical engineering applications in their ability to exploit the underlying structure of the input data for both decoding and visualization purposes [2]. Biomedical classification problems typically deal with multi-dimensional signals, in time, space, and/or frequency domains. In deep learning, convolutional neural networks (conv nets) have been evaluated for the classification of event-related potentials (ERPs), steady-state visual evoked potentials, and motor imagery [3], [4]. Typically, the underlying structure of the input signal is split between its two main dimensions: the temporal dimension that represents the different time points, and the spatial dimension that represents the different sensors that are used during EEG recording. The choice of a deep learning architecture can take advantage of the knowledge about the problem, shifting the human work from feature engineering to architecture engineering.

In EEG, the position of the sensors respects a system, such as the International 1020 system [5]. The number of sensors has increased over time, from 10-20 to up 5% system for high density EEG [6], [7]. Such high density system is useful for topographic methods and tomographic signal source localization methods [8]. Contrary to computer vision problems where there exist now large labeled datasets of images, EEG signal processing has some disadvantages: the number of trials available for an individual is often low, the EEG signal has a low signal-to-noise ratio, and ERPs based paradigms provide an unbalanced class distribution. Conv nets have been used for the ERP detection in tasks such as the P300 speller [9] and rapid serial visual presentation tasks [10], [11]. These architectures include a first convolutional layer with the purpose of achieving spatial filtering while the next layers provide some temporal filtering. Some architectures used on EEG signal are simple adaptation of architectures that are used in computer vision. While the convolution aims at representing the behavior of receptive fields in the visual cortex, the semantic of the convolutions in the EEG signal must represent meaningful transformations in the signal. If a 2D convolution is applied on the signal and the size in the spatial dimension is inferior to the size of the dimension, then such a convolution will be biased in relation to the order of the channels in the input matrix. Such a choice may connect channels that are not close from each other on the scalp. While ERP analysis can be achieved only in the spatial and time domains, specific architectures have been proposed for the detection of steady-state visual evoked potentials, where spatial filtering was achieved through a convolutional layer and the units in the time domain were normalized in the Fourier domain to extract high level features [12]. In addition to the architecture, batch normalization, the choice of the activation functions, and optimization function have an impact on the performance. Because linear discriminant analysis provides efficient results for ERP single trial detection, using a fully connected hidden layer before the output layer is not necessary, and therefore the addition of the dropout feature may not be relevant. The contributions of this paper are related to the evaluation of new architectures for convolutional neural networks by comparing 2D and 3D convolution layers, and by proposing a remapping of the 1D spatial dimension into a 2D map. In the present study, we consider a dataset of EEG signals corresponding to event-related potentials when subjects participated to a go-nogo task, i.e. target detection task. The EEG signal was recorded with 64 channels in the 10-20 system. The paper is organized as follows: the 3D convolution on EEG signal and the proposed conv net architectures are presented

Department of Computer Science, College of Science and Mathematics, Fresno State University, Fresno, CA, USA. hcecotti@csufresno.edu

978-1-5386-1311-5/19/$31.00 ©2019 IEEE



(a) Order of the sensors in 1D

-	-	-	-	1	33	34	-	-	-	-
-	-	-	2	3	37	36	35	-	-	-
-	7	6	5	4	38	39	40	41	42	-
-	8	9	10	11	47	46	45	44	43	-
-	15	14	13	12	48	49	50	51	52	-
-	16	17	18	19	32	56	55	54	53	-
24	23	22	21	20	31	57	58	59	60	61
-	-	-	25	26	30	63	62	-	-	-
-	-	-	-	27	29	64	-	-	-	-
-	-	-	-	-	28	-	-	-	-	-

(b) 2D layout

Fig. 1. Biosemi cap layout with sensors position and order in 1D, corresponding 2D Layout for 64 sensors.

in Section II. The performances of these architectures are analyzed in Section III. Finally, the results are discussed in Section IV.

II. Methods

A. Layout system

In this study, we consider 64 sensors from the 10/20 layout system in relation to the order of the sensors given by BIOSEMI caps. Such a layout is depicted in Fig. 1. The numbers in the table on the left side correspond to the order of the sensors when they are placed in a vector. The corresponding layout in the 10-20 system is given in Fig. I. In this 2D representation, a large number of cells are empty, it is the case for the first and last row. In addition, the first and last column contain only a single element (P9 and P10). These elements can be placed in positions (8,2) and (8,10), respectively, or to be removed in order to reduce the size of the 2D space. We assume that the first row (FP1, FPz, and FP2) and the last row (Iz) do not contain any meaningful information for the classification of ERPs. This selection leads to a size of 8 $\times$ 9 for the spatial projection. The missing cells are filled by using a combination of the 4-neighborhood (Von Neumann neighborhood). For instance the cell at position $P(2,3)$ is estimated as $(P(2,4)+P(3,2))/2$.

B. 2D and 3D Convolutions

In a regular 2D convolution, the transformation is applied on the 2D feature maps in order to compute features from the 2D dimensions only. When it is applied to 3D data, we wish to extract patterns that are encoded and specific to each dimension. Therefore, it seems more judicious to perform 3D convolutions in the convolution stages of conv nets to extract discriminant features from each dimension [13]. For EEG signal, it represents 2 dimensions in the spatial domain, and 1 temporal dimension. For a regular 2D convolution with a mask of size $W_1 \times W_2$, the value of the unit $y(x_1,x_2)$ at the m^{th} feature map in the l^{th} layer is given by:

$$\begin{aligned} y^{m,l}_{x_1,x_2} &= f(\sigma^{m,l}_{x_1,x_2}) + w^{m,l}_0 \\ \sigma^{m,l}_{x_1,x_2} &= \sum_{m0=0,i1=0,i2=0}^{M,W_1,W_2} w^{m,l,m0}_{i1,i2} \cdot y^{m0,l-1}_{x_1+i1,x_2+i2} \end{aligned} \tag{1}$$

where f is an activation function (e.g. tanh, rectified linear unit function [14]). The set of weights $w^{m,l,m0}$ represents the connections between the unit at coordinate (x_1,x_2) at the m^{th} feature map in the l^{th} layer and its corresponding unit at the $m0^{th}$ feature map in the $(l-1)^{th}$ layer (the previous layer). $w^{m,l}_0$ is a threshold. The 3D convolution is performed by convolving a 3D mask to the cuboid of size $W_1 \times W_2 \times W_3$, formed by concatenating several windows in the spatial dimensions, along the temporal dimension. With such a process, the feature maps this 3D convolutional layer are linked to multiple windows in the previous layer to capture patterns that are unique in the 3 dimensions. The value of the unit $y(x_1,x_2,x_3)$ at the mth feature map in the l^{th} layer is expressed by:

$$\begin{aligned} y^{m,l}_{x_1,x_2,x_3} &= f(\sigma^{m,l}_{x_1,x_2,x_3} + w^{m,l}_0) \\ \sigma^{m,l}_{x_1,x_2,x_3} &= \sum_{\substack{m0=0,\\ i1=0,\\ i2=0,\\ i3=0}}^{\substack{M,W_1,\\ W_2,W_3}} w^{m,l,m0}_{i1,i2,i3} \cdot y^{m0,l-1}_{x_1+i1,x_2+i2,x_3+i3} \end{aligned} \tag{2}$$

The set of weights $w^{m,l,m0}$ represents the connections between the unit at coordinate (x_1,x_2,x_3) at the m^{th} feature map in the l^{th} layer and its corresponding unit at the $m0^{th}$ feature map in the $(l-1)^{th}$ layer. The number of feature maps corresponds to the number of feature types that are extracted in the input(s) cuboid(s).

C. Database

The EEG database was previously used in [15], [16]. The experimental protocol is briefly detailed: 16 participants viewed a series of simulated images from a desert metropolitan environment using a rapid serial visual presentation (RSVP) paradigm containing 110 target images (scene with a person holding a gun) and 1346 non-target images (a scene without any people). The EEG signals were sampled at 1024 Hz from 64 scalp electrodes arranged in a 10-20 montage using a BioSemi Active Two system. The signal was bandpassed [0.1-21.33 Hz]with a 4th order

TABLE I

2D LAYOUT REPRESENTATION FOR 64 SENSORS IN THE 10-20 SYSTEM.

	1	2	3	4	5	6	7	8	9	10	11
1	-	-	-	-	FP1	FPz	FP2	-	-	-	-
2	-	-	-	AF7	AF3	AFz	AF4	AF8	-	-	-
3	-	F7	F5	F3	F1	Fz	F2	F4	F6	F8	-
4	-	FT7	FC5	FC3	FC1	FCz	FC2	FC4	FC6	FT8	-
5	-	T7	C5	C3	C1	Cz	C2	C4	C6	T8	-
6	-	TP7	CP5	CP3	CP1	CPz	CP2	CP4	CP6	TP8	-
7	P9	P7	P5	P3	P1	Pz	P2	P4	P6	P8	P10
8	-	-	-	PO7	PO3	POz	PO4	PO8	-	-	-
9	-	-	-	-	O1	Oz	O2	-	-	-	-
10	-	-	-	-	-	Iz	-	-	-	-	-

Butterworth filter, then downsampled to 64 Hz. The epoching was performed with 51 time points (800 ms) post stimulus. The input EEG data are stored into a 4D structure of size $N_{s1} \times N_{s2} \times N_t \times N_{ex}$ where $N_{s1} = 8$, $N_{s2} = 9$ represent the spatial dimension, $N_t = 51$, and N_{ex} corresponds to the total number of epochs.

D. Architectures

The goal is to classify the brain evoked responses corresponding to target and non-target images. We compare 3 main conv net architectures that are presented in Table II. The first column represents the size of each feature map, the second column corresponds to the number of feature maps, finally the type is the function that links the layer in the row to its previous layer. For a 3D convolution $(3, 3, 3)$ it means that $(W_1 = 3, W_2 = 3, W_3, 3)$. The relu activation function is used in all the layers, and a softmax function is used before reaching the final outputs, the training is performed with the Adam optimizer algorithm. Each architecture is evaluated with 2 sets of parameters corresponding to the number of feature maps. The first architecture includes 3 3D convolutional layers, the second architecture has 4 3D convolutional layers, while the third architecture has 2 3D convolutional layers. It is worth noting that the mask size of the first conv layer of the third architecture has the maximum dimension in the first 2 components. Thus, it corresponds to a 2D convolution, as if the 2 first dimensions would be the same, and the next layer has a unique dimension (in time). The second conv layer is equivalent to a 1D convolution as there is only 1 dimension. In the next section, we assess the performance with the area under the ROC curve by evaluating a model for each subject of the dataset using a 5-fold cross validation.

III. RESULTS

The AUC for the different architectures (A1 to A6) are presented in Table III. The best performance was achieved with the first architecture (A1) with a mean AUC across subjects of 928± 50. Using a Wilcoxon signed-rank test for pairwise comparisons, the results indicate that the number of feature maps has a significant impact on the performance: architectures A1, A3, A5, provide a better performance than A2, A4, and A6, respectively ($p < 10e - 4$). Moreover, the results reveal that A1 provides the best performance compared to the other architectures ($p < 10e - 4$).

TABLE II

CONV NET ARCHITECTURES.

	Size	# feature maps		Type
Input	(9,8,51)	1	1	-
Output	(1,1,2)	1	1	Fully connected
		A1	A2	
Layer 1	(7,6,49)	4	2	Convolution 3D (3,3,3)
Layer 2	(5,4,47)	8	4	Convolution 3D (3,3,3)
Layer 3	(1,1,45)	16	8	Convolution 3D (5,4,3)
		A3	A4	
Layer 1	(7,6,49)	4	2	Convolution 3D (3,3,3)
Layer 2	(5,4,47)	8	4	Convolution 3D (3,3,3)
Layer 3	(3,2,45)	16	8	Convolution 3D (3,3,3)
Layer 4	(1,1,43)	32	16	Convolution 3D (3,2,3)
		A5	A6	
Layer 1	(1,1,49)	4	2	Convolution 3D (9,8,3) (2D)
Layer 2	(1,1,45)	16	8	Convolution 3D (1,1,5) (1D)

IV. DISCUSSION AND CONCLUSION

Knowledge about the underlying structure of the data to be classified provides key information about the design of the architecture to set up in convolutional neural networks. Such information has been exploited in computer vision and other problems where there exists an underlying structure in the data, i.e where there is the notion of neighborhood and relationships within the features. While the spatial domain in the EEG signal is typically represented in a single dimension, it corresponds to 2 dimensions from the surface of the scalp. Such information can be exploited to reduce the number of inputs in the units of a feedforward artificial network using convolution layers.

In this study, we have investigated the impact of 3D convolutions on ERP single-trial detection with 64 channels. From 64 channels, by removing 6 channels, we created a 2D projection of 72 (9*8) channels, which is superior to the original size in 1D. However, such a projection allows to process the signal at a finer scale through a deeper architecture by using small masks (e.g. $3 \times 3 \times 3$). With a regular 2D convolution in the spatial domain, the number of input units is 64, while with a 3D convolution, it is possible to filter the signal in both the spatial and temporal domains with 27 input units. Such a filter allows to decrease the number of input units while performing two actions at the same time: in the space and time domain. By using regular 2D convolutions, the signal is directly projected in a single dimension (in the time domain). The performance

TABLE III

PERFORMANCE FOR THE DIFFERENT ARCHITECTURES (AUC × 10E3).

Subject	3D convolution		3D convolution		2D+1D convolution	
	A1	A2	A3	A4	A5	A6
1	921± 22	900±10	871±15	865±27	903±14	888±11
2	937± 24	918±22	899±32	886±30	922±23	902±20
3	947± 24	927±23	917±37	908±40	936±28	916±26
4	934± 31	916±28	911±36	895±42	924±32	903±34
5	936± 28	916±25	911±32	893±38	921±30	902±31
6	933± 27	914±24	905±33	890±36	919±28	898±31
7	938± 28	919±26	910±33	897±37	925±30	903±32
8	912± 78	894±72	885±75	872±77	897±80	873±91
9	917± 75	901±71	892±74	881±76	905±78	882±90
10	917± 72	901±67	893±70	882±73	904±74	880±86
11	920± 69	903±65	895±67	885±70	907±72	884±83
12	924± 67	909±65	901±67	891±70	912±70	890±82
13	926± 65	912±63	905±66	895±69	916±69	895±81
14	930± 64	915±62	908±65	899±68	919±67	896±78
15	928± 62	914±61	905±63	898±66	917±66	893±77
16	927± 60	914±59	903±62	896±64	915±64	892±75
mean	928±50	911±46	901±52	890±55	915± 52	893±58

comparison across 16 subjects has validated the approach of 3D convolutions over 2D convolutions. In addition to these encouraging results with such an architecture on 64 channels, the impact of 3D convolutions would be more beneficial by using magnetoencephalography (MEG) that can have up to 306 channels, and high density EEG, with up to 256 channels. In these cases, a single dimension can represent a challenge in terms of the number of features, and extracting features along such a high dimensional space can be difficult when using a low number of examples. A deep 3D architecture would leverage such problems. Furthermore, high density EEG recordings have been shown to improve localization of epileptic foci in surgical candidates compared to visual interpretation of the conventional scalp EEG [17]. Hence, 3D convolutions would be a suitable approach for such clinical applications. Finally, the EEG signal remains what happens at the surface of the scalp. The proposed projection could be improved and updated in relation to the estimation of the empty cells in 2D, a key question to address is what are the ideal values to be placed in those cells, if they have to be estimated in relation to the neighborhood, or if they should remain empty with a dummy value. The extraction of sources would provide information in 3D dimensions (x,y,z), hence it would be possible to extract 4D (x,y,z,time) features from the EEG signal for the classification using a 4D structure as input. Future works will validate the proposed approach on high density EEG and MEG to extract local spatio-temporal features that have the potential to improve single-trial performance.

REFERENCES

[1] K.-R. Müller, M. Tangermann, G. Dornhege, M. Krauledat, G. Curio, and B. Blankertz, "Machine learning for real-time single-trial EEG-analysis: From braincomputer interfacing to mental state monitoring," *J. of Neuroscience Methods*, vol. 167, pp. 82–90, 2008.

[2] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, pp. 5391–5420, Nov. 2017.

[3] N. Lu, T. Li, X. Ren, and H. Miao, "A deep learning scheme for motor imagery classification based on restricted boltzmann machines," *IEEE Trans. Neural Sys. Rehab. Eng.*, vol. 25, no. 6, pp. 566–576, Jun. 2017.

[4] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *J Neural Eng.*, vol. 14, no. 1, p. 016003, Feb. 2017.

[5] G. E. Chatrian, E. Lettich, and P. L. Nelson, "Ten percent electrode system for topographic studies of spontaneous and evoked EEG activity," *Am. J. EEG Technol.*, vol. 25, pp. 83–92, 1985.

[6] R. Oostenveld and P. Praamstra, "The five percent electrode system for high-resolution EEG and ERP measurements," *Clinical Neurophysiology*, vol. 112, pp. 713–719, 2001.

[7] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *Neuroimage*, vol. 34, pp. 1600–1611, 2007.

[8] C. J. Chu, "High density EEG what do we have to lose?" *Clin Neurophysiol.*, vol. 126, no. 3, pp. 433–434, Mar. 2015.

[9] H. Cecotti and A. Gräser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 433–445, Mar. 2011.

[10] H. Cecotti, M. P. Eckstein, and B. Giesbrecht, "Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering," *IEEE Trans. Neural Networks and Learning Systems*, vol. 15, pp. 2030–42, Nov. 2014.

[11] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces," *arXiv*, pp. 1–30, 2018.

[12] H. Cecotti, "A time-frequency convolutional neural network for the offline classification of steady-state visual evoked potential responses," *Pattern Recognition Letters*, vol. 32, no. 8, pp. 1145–1153, 2011.

[13] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[14] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Proc. of the 12th Int. Conf. on Computer Vision (ICCV'09)*, 2009, pp. 2146–2153.

[15] A. R. Marathe, A. J. Ries, V. J. Lawhern, B. J. Lance, J. Touryan, K. McDowell, and H. Cecotti, "The effect of target and non-target similarity on neural classification performance: a boost from confidence," *Frontiers in Neuroscience*, vol. 9, pp. 1–11, 2015.

[16] H. Cecotti, A. Marathe, and A. J. Ries, "Optimization of single-trial detection of event-related potentials through artificial trials," *IEEE trans. Biomed. Eng.*, pp. 1–7, 2015.

[17] V. Brodbeck, A. M. Lascano, L. Spinelli, M. Seeck, and C. M. Michel, "Accuracy of EEG source imaging of epileptic spikes in patients with large brain lesions," *Clin Neurophysiol.*, vol. 120, no. 4, pp. 679–85, Apr. 2009.